Interpretable and Explainable AI for

Prof. K.P (Suba) Subbalakshmi, FNAI
Dept. of E.C.E, Stevens Institute of Technology
Jefferson Science Fellow
ksubbala@stevens.edu
Https://www.kpsuba.com
About the PI

- Prof. with over 21 years of experience
- Fellow, National Academy of Inventors, 2018
- NJ Inventors Hall of Fame Award for work on deception detection from text
- NASEM’s Science and Technology Experts Group for ODNI

Interpretable and Explainable NLP

Fake News Detection

Gender identification

Authorship Similarity Detection

Are they written by the same author or not?
### Example Explanations

<table>
<thead>
<tr>
<th>Label</th>
<th>Source Statement</th>
<th>Top 3 Tweet Statement</th>
<th>Attention Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Witness: Police allegedly stopped Mike Brown after yelling at him to walk on sidewalk. Ferguson <a href="http://t.co/XG00R6w0k6">http://t.co/XG00R6w0k6</a></td>
<td>@Agent Kindi @SecretService The SecretService Protects Obama PresidentObama He Get’s Threats All The Time. @MichaelSkolnik @Supreme Power @MichaelSkolnik You so edgy. @TimmyTurnUp @MichaelSkolnik @Supreme Power U just want to say ”white is guilty, because they white”? In Moscow black guys sold drugs...</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table 2: Top three tweets (based on attention values) for the Ferguson event in the PHEME dataset. Label corresponds to the ground truth and a label value of 1 indicates fake news. This tweet was classified correctly by the proposed model.

**Tweet level explanations**

**Content feature level explanations**
## Example Results

<table>
<thead>
<tr>
<th>Goal</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authorship identification</td>
<td>Over 90% accuracy</td>
</tr>
<tr>
<td>Fake News Detection</td>
<td>Beats SOTA performance while also providing layers of explanation</td>
</tr>
<tr>
<td>Gender identification</td>
<td>Accuracy between 75% and 85%</td>
</tr>
</tbody>
</table>
Sample Relevant Publications and Patents


• Mingxuan Chen, Xinqiao Chu and K.P. Subbalakshmi, "MMCoVaR: Multimodal COVID-19 Vaccine Focused Data Repository for Fake News Detection and a Baseline Architecture for Classification", ASONAM 2021


• More at https://www.kpsuba.com